DOCUMENT RESUME

ED 189 995 HE 012 949

AUTHOR Kaufman, Brian J.; Madden, Joseph M.
TITLE The Development of Behaviorally Anchored Rating Scales for Student Evaluation of College Teaching Effectiveness.
PUB DATE Apr 80
NOTE 10p.; Paper presented at the Annual Meeting of the Eastern Psychological Association (51st, Hartford, CT, April 9-12, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Behavior Rating Scales; *College Faculty; *College Instruction; Evaluation Criteria; *Faculty Evaluation; Higher Education; Item Analysis; Material Development; *Student Evaluation of Teacher Performance; Teacher Behavior; *Teacher Effectiveness

ABSTRACT

A study was conducted to develop a behaviorally anchored rating scale for the evaluation of college teaching that could be used in all academic areas and that would reduce leniency, central tendency, and other rater biases. The Smith-Kendall procedure was modified by having undergraduate subjects generate behavioral examples for specified levels of effectiveness. A Likert-type scale was included to provide a frame of reference to aid judges in the generation of mid-range behaviors. For nine dimensions of teaching effectiveness, 243 critical behaviors were generated, 165 of which were eliminated because they did not meet criteria for retention as anchors. The seven scales that resulted encompassed dimensions of teacher effectiveness that students were best able to rate reliably. It is claimed that the high degree of consensus among judges, as well as the criteria for item retention, ensure that the final anchors are unambiguous, but generalizable to any classroom situation. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189995

The development of behaviorally anchored rating scales for student evaluation of college teaching effectiveness [1]

Brian J. Kaufman

and

Joseph M. Madden

Rensselaer Polytechnic Institute

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Brian J. Kaufman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

[1] Paper presented at the 51st annual meeting of the Eastern Psychological Association, April 9-12, 1980; Hartford, Connecticut.

HE 012 949

The Development of Behaviorally Anchored Rating Scales for Student Evaluation of College Teaching Effectiveness

As many as eighty-four percent of colleges and universities which participated in a recent survey reported using some form of standardized student evaluation of faculty. In addition to being the most influential factor in determining teaching effectiveness, this source of information accounted for between ten and forty percent of the weight that determines tenure and promotion decisions (Muse, 1979; Rippey, 1975). The use of these evaluations as a measure of teaching effectiveness attracted widespread attention as part of the "student revolt" of the last decade, when curriculum relevance, tenure decisions, and academic policies were objects of student unrest (Kent, 1966). Student evaluations are again attracting public interest because of the need to cut and account for spending at institutions of higher education.

The rating forms which are now used vary widely between and within institutions, ranging from the multiple choice descriptive format, to graphic scales and open-ended essay forms. Reviews of the literature show that student ratings are reliable, valid, and consistent, and as such may provide useful feedback for faculty, students, and administrators. The results of teacher evaluations appear in student course handbooks, budget proposals, and chairmens' reports to the dean. The methods of survey development also vary greatly, and many of the resulting forms are unorganized, ask too many questions, or overgeneralize the phrasing of survey items (Costin, Greenough, and Menges, 1971; Feldman, 1977;

Kent, 1966; McKeachie, 1969). These methods may result in less than desirable effectiveness as measures of performance. A method for performance evaluation or appraisal which would reduce these weaknesses is proposed by Smith and Kendall (1963) for the development of behaviorally anchored rating scales. This method uses independent groups for the development of dimensions and items of performance criteria. This procedure has several advantages: (a) members of the rater population construct the scales; (b) raters arrive at a consensus as to the construct validity, independence, and exhaustiveness of the dimensions; (c) the raters' own terminology is used to describe and define the specific behavioral examples, thereby eliminating gross performance descriptions and many response biases; and (d) in actual use, the raters document ratings with specific incidents they have observed. These factors would facilitate honest and conscientious ratings. Harari and Zedeck (1973) attempted to reduce the ambiguity, verbosity, and vagueness of student rating forms by developing a specialized behaviorally anchored rating scale for the evaluation of psychology professors only. The purpose of the present study is the development of a behaviorally anchored rating scale for the evaluation of college teaching which could be used in all academic areas and which would reduce leniency, central tendency, and other rater biases as described in Campbell, Dunnette, Arvey, and Hellervik (1973).

Method

Subjects. The subjects were 150 undergraduate coeds enrolled in five summer session psychology courses at a college in Brooklyn, and 450 undergraduates (about ten percent female) enrolled in engineering and humanities courses at an upstate New York private university. All subjects were participating on a voluntary basis.

Procedure. The scale development consisted of three iterative stages, whereby results obtained from one group became the stimuli for the next. The number of subjects differed for different stages of development. Subjects received an explanation of the methodology and goals of the research before participating in any of the stages.

Stage 1. The first stage utilized a modification of the original Smith-Kendall procedure for the generation of critical behavioral incidents. Nine student-generated dimensions of teaching behavior were identified through a rank-ordering procedure in the initial stages of research, as well as their definitions retaining student terminology. The dimensions-- Stimulation of Interest, Communication Skill, Accessibility, Responsiveness, Fairness, Grading Accuracy, Knowledge of Subject Material, Preparation, and Patience-- and their definitions were arranged on survey forms, such that each subject was presented with two dimensions for consideration. A ten-point Likert-type scale ranging from 1 (none) to 10 (extreme amount) appeared below each dimension. Subjects (n= 150) were students enrolled in at least one upper level psychology course at a college in Brooklyn, who were instructed to list behavioral examples for levels 2, 4, 6, 8, and 10 of each of the dimensions. The purpose of this method was an attempt to reduce the problem raters have in generating examples of behaviors representative of average, or non-critical performance (Harari & Zedeck, 1973; Landy & Farr, 1980; Landy & Guion, 1970; Smith & Kendall, 1963). Fifty-three usable surveys were returned, generating 243 behavioral examples. A usable survey was one which had at least four of the five levels defined by behaviors for a teaching dimension.

Stage 2. Subjects (n=380) were upper level students who were enrolled in a wide range of engineering and humanities coursed at an upstate New York private university. The behavioral examples generated in Stage 1 were

randomized using a Fortran procedure and were distributed across five survey forms, such that wach subject received approximately 48 behaviors to judge. This was done to make the task less formidable, minimizing subject mortality and completion time. A reference page, composed of the nine teaching dimensions and definitions, was distibuted with the instructions to "match each behavioral example to the dimension it best represents" (a procedure which Smith & Kendall, 1963, call "retranslation"). Elimiation of incomplete forms resulted in 157 surveys for inclusion in the data analysis. To be retained for consideration in Stage 3, an item had to be assigned to the same dimension by at least seventy-five percent of the judges, with not more than twenty percent of the judges assigning it to any other single dimension. Any ambiguous behavioral examples were unable to meet these criteria of rater consensus and were eliminated. Analysis of item allocation indicated that judges were unable to differentiate the dimension "Accessibility" from "Responsiveness", nor "Knowledge" from "Preparation"; hence, these pairs of dimensions were combined to yield seven dimensions instead of the original nine. After the retranslation process, 78 items were retained tor analysis in Stage 3.

Stage 3. Assignment of scale values to anchor the behavioral examples consisted of grouping the retained items by dimension and placing them in random order below the corresponding dimension definition. A 10-point Likert-type scale anchored only with 1 (least) to 10 (most) was included below the definition. Subjects (n=71) were instructed to rate each item on the degree of effectiveness that it represented, relevent to the performance dimension with whoch it was grouped. All behavioral examples were reworded from actual behaviors to expected behaviors, according to rationale explained by Smith and Kendall (1963). The rewording is believed to encourage

conscientiousness of raters because the anchors (a) become so concrete that in view of the previous consensus by the peer group, central tendency effects should be minimized, and (b) are so verifiable that the raters' judgement, insight, and values are later potentially challenged if subsequent ratee behavior would fail to confirm the raters' prediction. Criteria for anchor value retention was a standard deviation value of less than 1.5.

Results

Stage 1 resulted in the generation of 243 critical behaviors for nine independent dimensions of teaching effectiveness. The retranslation procedure of Stage 2 eliminated 165 behaviors which did not meet criteria for retention as anchors (senventy-five percent consensus among raters, with less than twenty percent allocation to any other single dimension). The dimension Responsiveness was combined with Accessibility, and the dimensions Knowledge of Subject Material and Preparation were combined after reviewing the examples and allocation patterns. The assignment of scale values to anchors in Stage 3 resulted in an average of six scale values anchored for each of the seven dimensions, with an average of nine behavioral anchors for each dimension.

Discussion

Modification of the Smith-Kendall procedure by having subjects generate behavioral examples for specified levels of effectiveness helped to reduce the problematic gap which usually occurs between the high and low ends of a behaviorally anchored rating scale. The inclusion of a Likert-type scale was intended to provide a frame of reference to aid judges in the generation of

mid-range behaviors. The mehtodology calls for the generation of anchors that will be meaningful to the raters who will be using them to rate someone; hence, a result of standard procedures for the development of behaviorally anchored rating sclaes is the exclusion of meaningless items. Landy and Guion (1970) and Harari and Zedeck (1973) note that meaningful stimuli will tend to polarize. The mid-range gap, then is indicative of the meaningfulness of the resulting anchors which are generated easily for the extreme ends of the scale.

Analysis of data in the retranslation process showed that raters were unable to differentiate between those behaviors intended to illustrate the dimensions of Knowledge of Subject Material and those of Preparation. It may be that the student judges in Stage 1, while understanding the importance of these dimensions of teaching, have little or no way of knowing how to delineate these dimensions ~~from~~ based on the students' observation of classroom behavior. For example, a teacher who is well versed in his field may not need to spend as much time in preparation for class as a teacher who is not as familiar with the subject material. The student would have no way of knowing the extent of the contribution that either dimension would have in accounting for the resulting behavior. In the same light, judges could not distinguish between the behaviors generated in Stage 1 for the dimensions Accessibility form those generated for Responsiveness. This apparent inability of judges to differentiate between behaviors representing these dimensions resulted in the combinations of Knowledge with Preparation and Accessibility with Responsiveness. The seven resulting scales encompass those dimensions of teacher effectiveness which students are best able to rate reliably. The high degree of consensus among judges, as well as the criteria for item retention,

ensure that the final anchors retained after Stage 3 are unambiguous, but generalizable to any classroom situation. The use of a large, diverse sample representative of several disciplines in the scale construction makes the scale useful for the comparison of student attitudes towards college teaching measured across more than one specific course or department.

References

Campbell, J.P., Dunnette, M.D., Arvey, R.D., & Hellervik, L.V. The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 1973, 57, 15-22.

Costin, F., Greenough, W.T., and Menges, R.J. Student ratings of college teaching: Reliability, validity, and usefulness. Review of Educational Research, 1971, 41, 511-535.

Feldman, K.A. Consistency and variability among college students in rating their teachers and courses: A review and analysis. Research in Higher Education, 1977, 6, 223-274.

Harari, O., & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology, 1973, 58, 261-265.

Kent, L. Student evaluation of teaching. Educational Record, 1966, Summer, 376-406.

Landy, F.J., & Farr, J.L. Performance rating. Psychological Bulletin, 1980, 87, 72-107.

Landy, F.J., & Guion, R.M. Development of scales for the measurement of work motivation. Organizational Behavior and Human Performance, 1970, 5, 93-103.

McKeachie, W.J. Student ratings of faculty. AAUP Bulletin, 1969, 55, 439-444.

Muse, C.J. Faculty evaluation. Improving College and University Teaching, 1979, 27, 43-46.